Davide Belli
11887532
davidebelli95@gmail.com

Homework Assignment 6
Machine Learning 3, 17/18

2018-05-13

In this Homework, I discussed possible interpretations with Gabriele Cesa.

1 Problem 1

1.1 Question a

Reminding that $q(z)$ is a distribution and $cq(z_0)$ represents the evaluation of z_0 through the function $cq(\cdot)$:

1: procedure RejectionSampler
2: n $\leftarrow 0$
3: $samples \leftarrow \emptyset$
4: while $n < Nsamples$ do
5: $z_0 \leftarrow$ sample($q(z)$)
6: $u_0 \leftarrow$ sample($Uniform([0, cq(z_0)])$)
7: if $y_0 < \tilde{p}(z_0)$ then
8: $samples \leftarrow samples \bigcup z_0$
9: $n \leftarrow n + 1$
return

1.2 Question b

The samples are independent since they are generated independently from the known distribution $q(z)$.

1.3 Question c

$$w_n = \frac{\tilde{r_n}}{\sum_m \tilde{r_m}} = \frac{\frac{\tilde{p}(z^{(n)})}{q(z^{(n)})}}{\sum_m \frac{\tilde{p}(z^{(m)})}{q(z^{(m)})}}$$

1.4 Question d

$$\begin{aligned}
A(x_{t+1}|x_t) &= \min\Big(1, \frac{\tilde{p}(x_{t+1})q(x_t|x_{t+1})}{\tilde{p}(x_t)q(x_{t+1}|x_t)}\Big) \\
&= \min\Big(1, \frac{\tilde{p}(x_{t+1})q(x_t|x_{t+1})}{\tilde{p}(x_t)q(x_{t+1})}\Big) \\
&= \min\Big(1, \frac{\tilde{p}(x_{t+1})q(x_t, x_{t+1})}{\tilde{p}(x_t)q(x_{t+1})^2}\Big) \\
&= \min\Big(1, \frac{\tilde{p}(x_{t+1})q(x_{t+1}|x_t)q(x_t)}{\tilde{p}(x_t)q(x_{t+1})^2}\Big) \\
&= \min\Big(1, \frac{\tilde{p}(x_{t+1})q(x_t)}{\tilde{p}(x_t)q(x_{t+1})}\Big) \\
&= \min\Big(1, \frac{p(x_{t+1})q(x_t)}{p(x_t)q(x_{t+1})}\Big)
\end{aligned}$$

1.5 Question e

They are not. Just by looking at the equation found in the previous answer, we see that the acceptance rate of the new sampled proposal point depends on x_t. However, the sampling itself is not dependent on x_t: $x_{t+1} \sim q(x|x_t) = q(x)$.

1.6 Question f

$[0.34, 0.34, 2.67, 0.82, 0.82]$

1.7 Question g

Rejection and Importance samplers poorly perform in high dimensionality. Regarding Rejection sampling, the rejection rate increases very fast (possibly exponentially) with the number of dimensions D. This is due to the fact that we want $q^{'}(z) = kq(z) \geq p(z) \forall z$, and with higher dimensions, the ratio of volume under p by the volume under q decreases very fast. For Importance sampling, a problem is that the number of terms in the summation $\mathbb{E}[f] \sim \sum^{L} p(z^l) f(z^l)$ grows exponentially. Also, a relevant part of the regions in space z will poorly contribute to that summation, because of a low value for p(z) or f(z). On the other hand, MC sampling methods (which Independence sampling belongs to), scales well with high dimensionalities. Although naive methods can be further improved to fight the curse of dimensionality, Monte Carlo sampling do not have to face the problems described for Rejection and Importance sampling. This is achieved by modeling the problem using Markov Chains, and making every sampling only dependent on the previous one, by introducing a constant matrix T describing transition probabilities between every node in the Chain.

2 Problem 2

To derive the Gibbs sampling for the posterior distribution it is enough to express separately the posteriors for the two parameters, μ and τ. Using the results in Bishop 2.3.6 to express the posterior of a Gaussian (knowing the mean and having a Gamma prior over the variance; knowing the variance and having a Gaussian prior over the mean) we can sample alternately from:

$$p(\tau|x,\mu) = \mathcal{N}\Big(\mu\Big|\Big(\frac{\tau^{-1}}{s_0+\tau^{-1}}\cdot\mu_0 + \frac{s_0}{s_0+\tau^{-1}}\cdot x\Big), \Big(s_0^{-1}+\tau\Big)^{-1}\Big)$$
$$p(\mu|x,\tau) = Gamma\Big(\tau\Big|a+\frac{1}{2}, b+\frac{1}{2}(x_n-\mu)^2\Big)$$

Reminding that $\tau^{-1} = \sigma^2$ and computing μ_{ML} over the single sample x, it reduces to that particular value

3 Problem 3

3.1 Question 1

Using:

$$p(z_{dn} = k|\vec{\theta}_d) = \theta_{dk}$$
$$p(w_{dn} = w|z_{dn} = k, \vec{\phi}) = \phi_{kw}$$
$$\prod_{n=1}^{N_d} p(z_{dn}|\vec{\theta}_d) = \prod_{k=1}^{K} \theta_{dk}^{A_{dk}}$$
$$\prod_{d=1}^{D}\prod_{n=1}^{N_d} p(w_{dn}|z_{dn},\vec{\phi}) = \prod_{k=1}^{K}\prod_{w} \phi_{kw}^{B_{kw}}$$

$$p(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\phi}|\beta,\alpha) = p(\boldsymbol{W}|\boldsymbol{\phi})p(\boldsymbol{\phi}|\beta)p(\boldsymbol{Z}|\boldsymbol{\Theta}), p(\boldsymbol{\Theta}|\alpha) =$$
$$= \Big(\prod^{K}\frac{1}{B(\beta)}\prod_{\boldsymbol{w}}\phi_{kw}^{\beta-1}\Big)\Big(\prod^{D}\frac{1}{B(\alpha)}\prod_{k}^{K}\theta_{dk}^{\alpha-1}\prod^{N}\theta_{dk}\phi_{kw_{dn}}\Big) =$$
$$= \frac{1}{\prod_d^D B(\alpha)\prod_k^K B(\beta)}\prod_k^K\Big(\prod_d^D\theta_{dk}^{A_{dk}+\alpha-1}\prod_{\boldsymbol{w}}\phi_{kw}^{B_{kw}+\beta-1}\Big)$$

3.2 Question 2

Noting that the content of the integrals resemble unnormalized Dirichlet distributions, the integrals can be evaluated with the normalizing Beta function.

$$\int p(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\phi}|\beta,\alpha)d\boldsymbol{\theta}_d d\boldsymbol{\phi}_k = \int \frac{1}{\prod_d^D B(\alpha)\prod_k^K B(\beta)}\prod_k^K\Big(\prod_d^D\theta_{dk}^{A_{dk}+\alpha-1}\prod_{\boldsymbol{w}}\phi_{kw}^{B_{kw}+\beta-1}\Big)d\boldsymbol{\theta}_d d\boldsymbol{\phi}_k =$$

$$= \frac{1}{\prod_d^D B(\alpha) \prod_k^K B(\beta)} \int \prod_k^K \prod_d^D \theta_{dk}^{A_{dk}+\alpha-1} d\boldsymbol{\theta}_d \int \prod_k^K \prod_{\boldsymbol{w}} \phi_{kw}^{B_{kw}+\beta-1} d\boldsymbol{\phi}_k =$$
$$= \prod_d^D \frac{B(A_{d1}+\alpha, ..., A_{dk}+\alpha)}{B(\alpha)} \prod_k^K \frac{B(B_{k1}+\beta, ..., B_{kw}+\beta)}{B(\beta)}$$

3.3 Question 3

The Gibbs sampling equation for the variable z_i can be derived from the integrated joint probability obtained in the previous step using:

$$p(\boldsymbol{W}, \boldsymbol{Z}|\beta, \alpha) = \prod_d^D \frac{B(A_{d1}+\alpha, ..., A_{dk}+\alpha)}{B(\alpha)} \prod_k^K \frac{B(B_{k1}+\beta, ..., B_{kw}+\beta)}{B(\beta)}$$
$$p(z_i|\boldsymbol{z}_{-i}, \boldsymbol{W}) = \frac{p(\boldsymbol{W}, \boldsymbol{Z})}{p(\boldsymbol{W}, \boldsymbol{Z}_{-i})}$$
$$= \frac{p(\boldsymbol{W}, \boldsymbol{Z})}{p(\boldsymbol{W}_{-i}, \boldsymbol{Z}_{-i}) p(w_i)}$$
$$\approx \prod_d^D \frac{B(A_{d,:}+\alpha)}{B(A_{d,-i}+\alpha)} \prod_k^K \frac{B(B_{k,:}+\beta)}{B(B_{k,-i}+\beta)} \frac{1}{\sum_{z_i} p(w_i|z_i) p(z_i)}$$

The factor in the last denominator can also be further developed by using again the integration over the joint probability for w_i, z_i including additional indicator functions to to only consider that particular document, word, topic.

4 Problem 4

4.1 Question a

$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$

4.2 Question b

$Cov(\boldsymbol{x}) = diag(\boldsymbol{\mu}^T(\mathbf{1} - \boldsymbol{\mu}))$

4.3 Question c

$\mathbb{E}[\boldsymbol{x}] = \sum^K \pi_k \boldsymbol{\mu}_k$

4.4 Question d

$$\log p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \log \prod^N \sum^K \pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k) =$$
$$= \sum^N \log \sum^K \pi_k \prod^D \mu_{ki}^{x_{in}} (1 - \mu_{ki})^{1-x_{in}}$$

4.5 Question e

Because there is no closed form solution for the optimization of the loglikelihood function, since it is a logarithm of summation, which is also difficult to compute analytically.

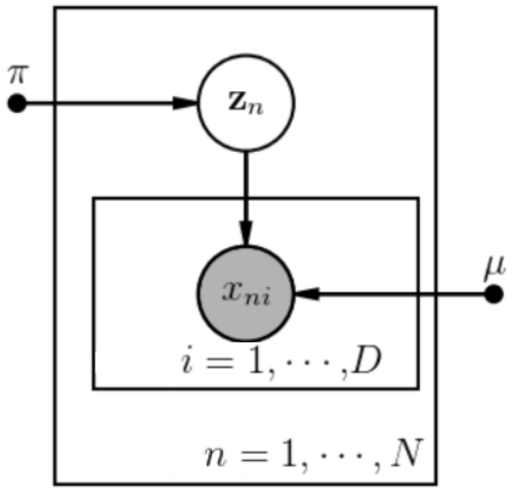
4.6 Question f

$$\log p(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \log \prod^N p(\boldsymbol{z}_n|\boldsymbol{\pi}) p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\mu}) =$$

$$= \sum^{N}\sum^{K}\Big(z_{nk}(\log \pi_k + \log p(\boldsymbol{x}_n|\boldsymbol{\mu}_k)\Big) =$$
$$= \sum^{N}\sum_{z_n}\Big(z_{nk}(\log \pi_{z_n} + \sum_{i}^{D} x_{in}\log \mu_{z_n i} + (1-x_{in})\log(1-\mu_{z_n i}))\Big)$$

4.7 Question g

Note that topics K can also be included in the plate notation, but for coherence with the notation used in Problem 3 were the iteration over K is not represented, I will omit that plate.



4.8 Question h

The lower bound we try to maximize is defined as:

$$\mathcal{B}(q_n(\boldsymbol{z}_n), \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum^{N}\sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n)\log\frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\mu}, \boldsymbol{\pi})}{q_n(\boldsymbol{z}_n)} =$$
$$= \sum^{N}\mathbb{E}_{q_n(\boldsymbol{z}_n)}[\log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\mu}, \boldsymbol{\pi})] + H(q_n(\boldsymbol{z}_n)) =$$
$$= \sum^{N}\sum_{\boldsymbol{z}_n}\Big(q_n(\boldsymbol{z}_n)(\log \pi_{z_n} + \sum_{i}^{D} x_{in}\log \mu_{z_n i} + (1-x_{in})\log(1-\mu_{z_n i})) + H(q_n(\boldsymbol{z}_n))\Big)$$

4.9 Question i

$$\tilde{\mathcal{B}}(q_n(\boldsymbol{z}_n), \boldsymbol{\mu}, \boldsymbol{\pi}) = \mathcal{B}(q_n(\boldsymbol{z}_n), \boldsymbol{\mu}, \boldsymbol{\pi}) + \sum^{N} H(q_n(\boldsymbol{z}_n)) + \lambda\Big(\sum_{i}^{K}\pi_i - 1\Big) + \sum_{N}\lambda_n\Big(\sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) - 1\Big)$$

The last part is just to force the all the distributions q_n to actually be distributions. Note that the distributions q_n are different for every $\boldsymbol{z}_n$.

4.10 Question j

In the E-step, we maximize the objective function with respect to $q_n(z_n)$:

$$0 = \frac{\partial}{\partial q_n(z_n)}\tilde{\mathcal{B}}(q_n(\boldsymbol{z}_n), \boldsymbol{\mu}, \boldsymbol{\pi})$$
$$0 = \log \pi_{z_n} + \sum_{i}^{D} x_{in}\log \mu_{z_n i} + (1-x_{in})\log(1-\mu_{z_n i}) - \log q_n(z_n) - 1 + \lambda_n$$

$$q_n(z_n) = \exp(\lambda_n - 1)\pi_{z_n} \prod_i^D \mu_{z_n i}^{x_{in}} (1 - \mu_{z_n i})^{1 - x_{in}}$$

After the E-step, the distributions $q_n(z_n)$ are the best approximators of the likelihood in order to raise the ELBO, ideally matching it with the likelihood level at that iteration. When this is accomplished, the KL-divergence between p and q is null, and $q_n(z_n)$ perfectly match the distributions p.

4.11 Question k

In the M-step, we optimize the objective function with respect to the parameters. For π:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \pi_k} \tilde{\mathcal{B}}(q_n(\boldsymbol{z}_n), \boldsymbol{\mu}, \boldsymbol{\pi}) \\
0 &= \frac{\sum^N \gamma(z_{nk})}{\pi_k} + \lambda \\
0 &= N_k + \lambda \pi_k \\
\lambda \sum^K \pi_k &= -\sum^K N_k \\
\lambda &= -N
\end{aligned}$$

$$\begin{aligned}
\pi_k &= -\frac{N_k}{\lambda} \\
\pi_k &= -\frac{N_k}{-N} \\
\pi_k &= \frac{N_k}{N}
\end{aligned}$$

5 Problem 5

Trivially, being this process symmetric,

$$\mathbb{E}[z^{(r)}] = 0$$

To find the second moment, we first note that this Random Walk can be represented with the following notation:

$$z^{(i+1)} = z^{(i)} + x^{(i)}$$

where

$$\begin{aligned}
p(x^{(i)} = 0) &= 0.5 \\
p(x^{(i)} = 1) &= 0.25 \\
p(x^{(i)} = -1) &= 0.25
\end{aligned}$$

By induction on i from r down to 0:

$$\begin{aligned}
z^{(r)} &= z^{(r-1)} + x^{(r)} \\
&= z^{(0)} + \sum_{i=1}^{r} x^{(i)} \\
&= 0 + \sum_{i=1}^{r} x^{(i)} \\
&= \sum_{i=1}^{r} x^{(i)}
\end{aligned}$$

Then, since every step $x^{(i)}$ is independent from all others and follows an identical distribution as described above,

$$x^{(i)} \perp\!\!\!\perp x^{(j)} \quad \forall i \neq j \implies \mathbb{E}[x^{(i)}, x^{(j)}] = 0$$

$$\begin{aligned}
\mathbb{E}\Big[(z^{(r)})^2\Big] &= \mathbb{E}\Big[(\sum_{i=1}^{r} x^{(i)})^2\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{r}\sum_{j=1}^{r} x^{(i)}x^{(j)}\Big] \\
&= \sum_{i=1}^{r}\sum_{j=1}^{r} \mathbb{E}\Big[x^{(i)}x^{(j)}\Big] \\
&= \sum_{i=1}^{r} \mathbb{E}\Big[x^{(i)}x^{(i)}\Big] \\
&= \sum_{i=1}^{r} \mathbb{E}\Big[(x^{(i)})^2\Big] \\
&= \sum_{i=1}^{r}\sum_{x^{(i)}} (x^{(i)})^2 p(x^{(i)}) \\
&= \sum_{i=1}^{r} \Big((-1)^2 * \frac{1}{4} + (0)^2 * \frac{1}{2} + (1)^2 * \frac{1}{4}\Big) \\
&= \sum_{i=1}^{r} \frac{1}{2} \\
&= \frac{r}{2}
\end{aligned}$$